Statistically Based Pattern Discovery Techniques for Biological Data Analysis

Andrew Hamilton-Wright\textsuperscript{1,3,3} and Daniel W. Stashuk\textsuperscript{4}

\textsuperscript{1} School of Rehabilitation Therapy, Queen’s University, Canada
\textsuperscript{2} Computing and Information Science, University of Guelph, Canada
\textsuperscript{3} Mathematics and Computer Science, Mount Allison University, Canada
\textsuperscript{4} Systems Design Engineering, University of Waterloo, Canada

\texttt{andrewhw@ieee.org}

\texttt{stashuk@pami.uwaterloo.ca}

Summary. A statistically based pattern discovery tool is presented that produces a rule-based description of complex data through the set of its statistically significant associations. The rules resulting from this analysis capture all the patterns observable within a data set for which a statistically sound rationale is available. The validity of such patterns recommends their use in cases where the rationale underlying a decision must be understood. High-risk decision making systems, a milieu familiar to many biologically-related problem domains, is the likely area of application for this technique. An analysis of the performance of this technique on a series of biologically relevant data distributions is presented, and the relative merits and weaknesses of this technique are discussed.

1.1 Introduction

Biological data is being collected into ever-larger databases, requiring strong analytic tools to help discover the important relationships encoded therein. Performing such discovery using statistically based analysis provides a simple and comprehensive understanding of the relationships discovered. There are, however, few general-purpose tools available that will uncover, with statistical rigour, the intricate and high-order relationships present in complex biological data and present the results in a form useful for classification or categorization of new data samples.

This chapter describes a classification tool that chooses labels based on rules derived only from statistically significant associations discovered in analysis of training data, providing a rule set that is insensitive to random noise in the training data, and is transparent: that is, they can be easily interpreted, explained, rationalized, confirmed, and made explicit, such that they can be considered verifiable acquired knowledge. In contrast, many other techniques do not concern themselves with statistical rigour, finding rules through the


\url{www.springerlink.com} © Springer-Verlag Berlin Heidelberg 2008
reduction of observed error; this leads both to an inability to generalize from training data and to being unable to understand how classifications are made.

Application of this tool to the classification of biological data provides a means of supporting clear and transparent decision making in high-risk domains. When computational support for high-risk decision making is provided, the transparency, explainability and statistical rigour of the decisions made has comparable importance to the overall accuracy of the system. In particular, transparency provides a means of incorporating multiple sub-decisions into an informed overall characterization while allowing all levels of the decision process to be understood by a human decision maker or stakeholder. The provision of such clarity rests on the quality of the patterns found, as well as on their ability to make characterizations.

The gathering of biological data is frequently quite expensive, both in terms of the time required for acquisition and in terms of the materials and equipment needed. It is therefore important that researchers be able to explore the data inter-relationships using tools that will locate informative patterns as well as provide good measures of the quality of the patterns found.

Broadly defined, algorithms that extract rules from training data through observation are referred to as “association classifiers” [1–12] and the extraction of such rules is termed “association mining” [13, 14]. Initially associated with an algorithm called “Apriori” used for mining databases [1], the term association classifier refers to an algorithm in which relationships discovered through analysis of training data are codified as rule-based patterns for use in classification. As our rules will make no assumptions regarding the relationships between features in the input data, we produce “generalized association rules” as described in [2], in contrast to the “pure” rules of [1].

Association classifiers are distinct from the more general class of supervised classification algorithms in that association classifiers deal explicitly with rule-based logic, while many other algorithms deal in metrics measuring a fit to a multi-dimensional error surface. To a party interested in the inference behind a particular classification, the rule-based approach provides a degree of transparency that is not available from distance-based inference.

The term “association mining” therefore encompasses many approaches, including such techniques as:

- tree generation algorithms [15, 16];
- extension matrix based techniques [17–21];
- fuzzy rule generation techniques, including:
  - fuzzy systems configured through evolutionary algorithms [22–25];
  - other neuro-fuzzy systems used for rule production [26–38];
- contingency table approaches, frequently based on rough sets [39–42]; and
- schemes using statistical clustering techniques to create input membership functions, combined with the generation of contingency tables [43–45].

All of these systems function by inspecting a training data set and producing a set of “rules” to be used in classification. The mechanism and quality of the rules produced may vary, however all these algorithms learn in a supervised